アスキーアート検索エンジンの実験的実装
An experimental implementation of a text art search engine

福井 悠 山田 晃嗣 小林 孝浩 関口 敦仁
FUKUI Yu YAMADA Koji KOBAYASHI Takahiro SEKIGUCHI Atsuhito

情報科学芸術大学院大学
Institute of Advanced Media Arts and Sciences

1. はじめに

テキストの電子化により、人々は文字集合から適切な文字を組み合わせて利用するようになった。その組み合わせや配置に空間的な意味を持たせたものがアスキーアートである。

アスキーアートは、文字情報しか受理することができないシステムの都合や、ネットワーク帯域の問題から静止画像による表現が制限された空間で利用されることがある。また文字情報と同様の手段で画像情報を送ることができるという手軽さから、あるいはアスキーアートというメディアがもつ魅力から、積極的にアスキーアートを選択する場合もある。

特に日本においては、かな漢字変換システムにアスキーアート辞書を標準的に搭載するようになっている。アスキーアートは日常的な表現手段のひとつといって問題ないだろう。

コンピュータネットワークの発達とともに、アスキーアートも爆発的に増加した。これまでにアスキーアートとそれをとりまく文化圏に関して、社会学的な考察は試みられているが、これを目的とした検索プログラムや機械的に分類しようという試みは確認できない。

本稿では、アスキーアートサーチエンジンの実験的実装について報告する。本実装は、あるアスキーアートと類似したアスキーアートの検索を直接の目的とするが、同時にアスキーアートの自動分類や特徴抽出、意味理解支援などの基盤技術としても応用可能である。

一部にはASCIIで符号化されたもののみをアスキーアートと呼び、ASCII以外で符号化されたものをテキストアートとして区別することもあるが、本稿ではそれらをまとめてアスキーアートとして扱う。

2. 本実装の概要と対象範囲

一般の画像検索システムと同様、アスキーアート検索エンジンにもメタデータ検索と例示画像検索の2つのタイプが考えらる。周辺もしくは内包されるキーワードからアスキーアートを検索するメタデータ検索であれば、最適化されているとはいえないものの、既存のサーチエンジンでも対応可能であろう。

本稿では、アスキーアートから類似のアスキーアートを検索するシステムを実装した。これは、一般の画像検索システムでいう例示画像検索にあたる。検索結果には完全に一致するアスキーアートだけではなく、検索対象と類似したアスキーアートも出力するようにした。また、検索結果には類似度を示すスコアを含めた。

実装はインデックス型とし、インデックスの生成にはN-gramを用いた。検索対象となるデータはあらかじめ分割され、ファイルの形で格納されているものとした。また、それぞれのアスキーアートは固定幅フォントを利用したものと仮定した。所謂半角文字と全角文字が混じることも想定していない。

本実装は、類似するアスキーアート間では同じ文字(もしくは連続する文字の組)が同程度出現しているという仮定にもとづいている。aalib[1]を用いるなど、静止画像から機械的な手法で生成されたアスキーアートに対して本実装は最適化されていない。

3. インデックス生成プログラム

実装言語として、Python 2.5を採用した。また、Pythonモジュール shelveを介しBerkeley DB 4.6をデータベースバックエンドとして利用している。

データベースの符号化文字集合にはUnicodeを用いた。なお、本稿ではU+からはじまる4桁の16進数によりUnicodeのコードポイントを示す。

ファイルに格納されたアスキーアートは、各行の文字数をもっとも長い行にあわせ、足りない部分をU+0000で埋めた。あわせてスペース(U+0020、U+3000など)をU+0000に置換、改行文字(一般的にはU+000A、U+000Dもしくはその組み合わせ)は削除した。このようにして正規化されたアスキーアートを対象にunigram、縦方向bigram、横方向bigramインデックスを生成、データベースにインデックスをキーと

しそれぞれのパターンの出現回数とファイル名を登録した。なお、インデックス生成時に文字数が足りない場合はU+0000を補った。

4. 検索プログラム

実装にはインデックス生成時と同様の言語を選択した。

検索対象のアスキーアートを前項と同様に正規化したのち、同様にunigram、横方向bigram、縦方向bigramに分解、それぞれのパターンごとに出現回数を集計した。分解されたパターンそれぞれについてデータベースに問い合わせ、出現回数の差分(の絶対値)を各ファイルごとに合計したものをスコアとした。

スコアが0の場合に類似度がもっとも高く、値が大きくなるにつれ一致度が低くなる。

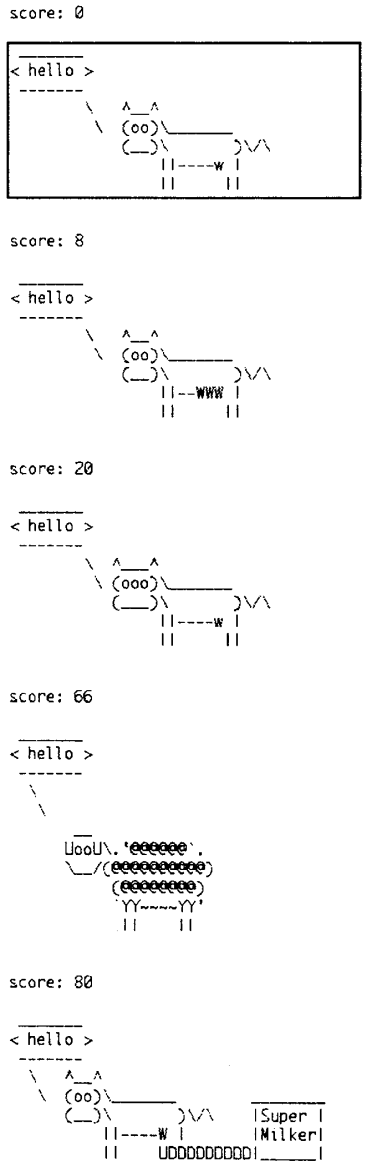

図1: 検索結果

5. 簡易評価

cawsay[2]に付属するcawファイルを利用してデータベースを構築、簡易的な評価をおこなった。実行結果を図1に示す。枠線で囲まれている部分が例示画像である。

実行結果は、おおむね感性評価と一致するものであると考えられる。しかし、データベースに登録されたアスキーアートの数が十分ではないことや、同一のパッケージにまとめられた(つまり、同じような画風をもち、同じような生成方法で作成された)アスキーアートのみを用いていることから、十分であるとは言い難い。

6. 今後の展開

本稿においてはサーチエンジンの一部であるインデックス生成プログラム及び検索プログラムの実験的実装を行った。本実装を正しく評価するには、ある程度の数があり、多様な画風を含んだアスキーアートが必要であると考えられる。必要なアスキーアートを収集するためのクローラ開発を当面の課題としたい。

実際にサービスとして機能させるためにはプロポーショナルフォントへの対応や、アスキーアートと自然言語が混じる文中におけるアスキーアートの切り出し及び適切なメタデータ付与などが必要であろう。

将来的には、アスキーアートの自動分類や分類されたアスキーアート群内での差異を比較することによる高位の意味解析、文脈に応じたアスキーアートの自動生成なども可能だろう。また、本稿においては言語処理的なアプローチを用いたが、扱う文字の拡大に伴い画像処理的な手法からのアプローチも必要になると考えられる。

参考文献
[1]aalib:
http://aa-project.sourceforge.net/aalib/
[2]cowsay:
http://www.nog.net/~tony/warez/cowsay.shtml